

D

Exhibit

We've been removing harmful content since YouTube started, but our investment in this work has accelerated in recent years.

By The YouTube Team  
Sep. 03. 2019

## The Four Rs of Responsibility, Part 1: Removing harmful content

INSIDE YOUTUBE

67

q

Official Blog  

1. Remove  
Over the past several years, we've redoubled our efforts to live up to our responsibility while preserving the power of an open platform. Our work has been organized around four principles:
  - Content that violates our policy as quickly as possible
  - Up alternative voices when people are looking for breaking news and information
  - Trusted, eligible creators and artists
2. Raise  
Content that violates our policy as quickly as possible
3. Reward  
Up alternative voices when people are looking for breaking news and information
4. Reduce  
the spread of content that brushes right up against our policy line

Before we do the work of removing content that violates our policies, we have to make sure the line between what we remove and what we allow is drawn in the right place — with a goal of preserving free expression, while also protecting and promoting a vibrant community. To that end, we have a dedicated policy development team that systematically reviews all of our policies to ensure they are current, keep our community safe, and do not stifle YouTube's openness.

After reviewing a policy, we often discover that fundamental changes aren't needed, but still uncover areas that are vague or confusing to the community. As a result, many updates are actually clarifications to our existing guidelines. For example, earlier this year we provided more detail about when we consider a "challenge" to be too dangerous for YouTube. Since 2018, we've made dozens of updates to our enforcement

Developing policies for a global platform

Over the next several months, we'll provide more detail on the work supporting each of these principles. This first installment will focus on "Remove." We've been removing harmful content since YouTube started, but our investment in this work has accelerated in recent years. Below is a snapshot of our most notable improvements since 2016. Because of this ongoing work, over the last 18 months we've reduced views on videos that are later removed for violating our policies by 80%, and we're continuously working to reduce this number further.<sup>1</sup>

guidelines, many of them minor clarifications but some more substantive. For particularly complex issues, we may spend several months developing a new policy. During this time we consult outside experts and YouTube creators to understand how our current policy is failing short, and consider regional differences to make sure proposed changes can be applied fairly around the world.

The hate speech update represented one such fundamental shift in our policies. We spent months carefully developing the policy and working with our teams to create the necessary trainings and tools required to enforce it. The policy was launched in early June, and as our teams review and remove more content in line with the new policy, our machine detection will improve in tandem. Though it can take months for us to ramp up enforcement of a new policy, the profound impact of our hate speech policy update is already evident in the data released in this quarter's Community Guidelines Enforcement Report.

The spikes in removal numbers are in part due to the removal of older comments, videos and channels that were previously permitted. In April 2019, we announced that we are also working to update our harassment policy, including creator-on-creator harassment. We'll share our progress on this work in the coming months.

we're investing significantly in these automated detection systems, and our engineering teams continue to update and improve them month by month. For example, an update to our spam detection systems in the second quarter of 2019 led to a more than 50% increase in the number of channels we terminated for violating our spam policies.

videos we removed in the second quarter of 2019 were first flagged by our automated systems.

human review to make nuanced decisions. Still, over 87% of the 9 million

categories are highly dependent on context and highlight the importance of can help to flag hate speech and other violate content, but these

that often looks the same — such as spam or adult content. Machines also

it's ever viewed. These systems are particularly effective at flagging content not exactly the same) to other content we've already removed, even before

is well-suited to detect patterns, which helps us to find content similar (but

potentially violate content and send it for human review. Machine learning

machines can catch at upload.

shared industry databases of hashes to increase the volume of content our

abuse images (CSAI) and terrorist recruitment videos, we contribute to

they are ever made available to view. For some content, like child sexual

(or "digital fingerprints") to catch copies of known violate content before

technology to flag content for our review teams. We sometimes use hashes

Once we've defined a policy, we rely on a combination of people and

## Using machines to flag bad content

Last week we updated our Community Guidelines Enforcement Report, a

same time period.

For example, the nearly 30,000 videos we removed for hate speech over the last month generated just 3% of the views that knitting videos did over the

We're determined to continue reducing exposure to videos that violate our policies. That's why, across Google, we've tasked over 10,000 people with detecting, reviewing, and removing content that violates our guidelines.

can become a larger issue.

We also recognize that the best way to quickly remove content is to anticipate problems before they emerge. In January of 2018 we launched our Intelligence Desk, a team that monitors the news, social media and user reports in order to detect new trends surrounding inappropriate content, and works to make sure our teams are prepared to address them before they

before they received a single view in the second quarter of 2019. We go to great lengths to make sure content that breaks our rules isn't widely viewed, or even viewed at all, before it's removed. As noted above, improvements in our automated flagging systems have helped us detect and review content even before it's flagged by our community, and consequently more than 80% of those auto-flagged videos were removed

**Removing content before it's widely viewed**

quarterly report that provides additional insight into the amount of content we remove from YouTube, why it was removed, and how it was first detected. That report demonstrates how technology deployed over the last several years has helped us to remove harmful content from YouTube more quickly than ever before. It also highlights how human expertise is still a critical component of our enforcement efforts, as we work to develop thoughtful policies, review content with care, and responsibly deploy our machine learning technology.

1 From January, 2018 - June, 2019

2 Nov 16, 2016, <https://youtube.googleblog.com/2016/11/more-parental-controls-available.html>

2 June 18, 2017, <https://www.blog.google/around-the-globe/google-europe/four-steps-we're-taking-today-fight-online-terror/>

2 Aug 1, 2017, <https://youtube.googleblog.com/2017/08/an-update-on-our-commitment-to-expanding-our-work-against-abuse-of.html>

2 July 31, 2017, <https://youtube.googleblog.com/2017/07/global-internet-forum-to-counter.html>

2 Dec 4, 2017, <https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of.html>

2 April 23, 2018, <https://youtube.googleblog.com/2018/04/more-information-faster-removals.html>

2 Dec 1, 2018, <https://youtube.googleblog.com/2019/06/an-update-on-our-efforts-to-protect.html>